On the Optimal Presentation Duration for Subjective Video Quality Assessment

Felix Mercer Moss, Ke Wang, Fan Zhang, *Member, IEEE*, Roland Baddeley, and David R. Bull, *Fellow, IEEE*

***Abstract*— Subjective quality assessment is an essential component of modern image and video processing for both the validation of objective metrics and the comparison of coding methods. However, the standard procedures used to collect data can be prohibitively time consuming. One way of increasing the efficiency of data collection is to reduce the duration of test sequences from a 10-s length currently used in most subjective video quality assessment (VQA) experiments. Here, we explore the impact of reducing sequence length upon perceptual accuracy when identifying compression artifacts. A group of four reference sequences, together with five levels of distortion, are used to compare the subjective ratings of viewers watching videos between 1.5 and 10 s long. We identify a smooth function indicating that accuracy increases linearly as the length of the sequences increases from 1.5 to 7 s. The accuracy of observers viewing 1.5-s sequences was significantly inferior to those viewing sequences of 5, 7, and 10 s. We argue that sequences between 5 and 10 s produce satisfactory levels of accuracy but the practical benefits of acquiring more data lead us to recommend the use of 5-s sequences for future VQA studies that use the double stimulus continuous quality scale methodology.**

***Index Terms*— DSCQS, HEVC, mean opinion scores, methodology, quality assessment, subjective testing, video databases, video presentation.**

I. Introduction

AS VISUAL display devices continue to pervade more of our lives, the need to provide their screens with optimized content is becoming no less critical. To ensure the qualitative experience of viewing digital video on these devices meets the increasing expectations of consumers, perceptual quality metrics are employed to evaluate how a sequence of images appears to the human visual system.

Automatic video quality assessment (VQA) is not a trivial challenge. The development of objective video quality metrics has accelerated in recent years, but despite the multitude of different algorithmic solutions available, few produce satisfactory correlations with ground truth data collected from human viewers [1]. To validate the performance of these objective measures, it is critical that researchers continue to provide new and diverse databases of test content, paired with reliable ground truth data obtained via subjective testing experiments.

Subjective testing is also important for the introduction of new video processing algorithms, such as those in compression or enhancement. For example, data from subjective quality assessment provide a perceptual benchmark to ensure that a new codec is providing significant performance gains over existing standards [2].

Improving the efficiency of the subjective testing process while preserving reliability is, therefore, a valuable area of research. The International Telecommunications Union (ITU) provide a set of canonical guidelines that aim to maintain a level of consistency in the methodologies used when collecting human data for such databases [3], [4]. While the ITU are to be applauded for encouraging a standardization of these practices, we believe the efficiency of one particular recommendation can be improved, while preserving the reliability of data they help collect.

The length of the test sequences used in a subjective experiment has significant practical implications upon the output of data. For example, for double stimulus (DS) methodologies, halving them in length could provide the researcher with the opportunity to collect the same amount of data in close to half the time or, in the same time, close to double the volume, dependent upon the length of voting time.[1] The ITU recommends using presentations of 10 s for moving pictures [3]; however, it is not clear how perceptual performance is affected if shorter sequences are used.

In addition to the associated practical benefits, there are both empirical and theoretical motivations for investigating the use of shorter sequence lengths for subjective testing. First, previous research indicates that observers become less critical of video presentation when clips are significantly longer than 10 s [5]. Second, shorter test sequences encourage more consistency in observers' viewing behavior [6], [7] and, consequently, their rating behavior. Third, the average shot length found in contemporary movies is significantly lower than 10 s [8].

To explore the impact of test sequence duration on opinion scores, four high-definition (HD) video sequences were

Manuscript received December 2, 2014; revised April 3, 2015 and May 13, 2015; accepted July 13, 2015. Date of publication July 29, 2015; date of current version October 27, 2016. This work was supported by the Engineering and Physical Sciences Research Council under Grant EP/J019291/1. This paper was recommended by Associate Editor P. Le Callet.

F. Mercer Moss, K. Wang, F. Zhang, and D. R. Bull are with the Department of Electrical and Electronic Engineering, University of Bristol, Bristol BS8 1UB, U.K. (e-mail: f.mercermoss@bristol.ac.uk; fan.zhang@bristol.ac.uk; dave.bull@bristol.ac.uk).

R. Baddeley is with the Department of Experimental Psychology, University of Bristol, Bristol BS8 1TU, U.K.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCSVT.2015.2461971

[1] Practically, most methodologies require additional voting time, reducing the magnitude of these savings. Assuming sequences are originally 10-s long and there has to be 5 s of additional voting time, DS methodologies could make 40% time savings (10 s + 10 s + 5 s = 25 s to 5 s + 5 s + 5 s = 15 s), while single stimulus (SS) methodologies could make 33% time savings (10 s + 5 s = 15 s to 5 s + 5 s = 10 s).

chosen along with five levels of distortion generated by High Efficiency Video Coding (HEVC) compression [9] and Gaussian blurring. Twenty-four human participants viewed and provided responses to five versions of each of these videos after being divided into five different sequence lengths ranging from 1.5 to 10 s. These data were used to address four specific research questions.

1) Does reducing sequence length below the recommended 10 s significantly affect the detection of distortion artifacts when using the double stimulus (DS) continuous quality scale (DSCQS) methodology?
2) Does significant variation in the level of distortion influence the strength of a potential duration effect?
3) Does the content of temporally consistent video sequences influence a potential duration effect?
4) Do observers feel more confident evaluating the video quality of certain sequence lengths over others?

This paper is structured as follows. Section II outlines some of the most influential subjective video databases and highlights some of the factors that can affect the results of human VQA experiments. The methodology of the main subjective experiment is outlined in Section III. The results are reported and discussed in Section IV, while the implications and future directions for this area of research are discussed in Section V.

II. Background

This section is divided into three subsections, the first of which introduces some of the most popular currently available subjective databases and how they are used. Section II-B explains some of the factors that can be responsible for both desired and undesired variation in opinion scores. Section II-C discusses previous research focused on test sequence length in subjective VQA experiments and how it might affect rating behavior.

A. Subjective Databases

Typically, subjective video databases include a set of original reference videos together with a group of variably distorted versions of each. All, or a smaller subset of these videos, are associated with a quality rating obtained by displaying the videos to human observers and recording their mean opinion scores (MOSs). Subjective video databases aim to provide researchers with a set of resources allowing them to assess the impact upon perception of variable spatial and temporal video quality. More specifically, they are primarily used for the validation of objective quality metrics or the performance comparison of different video coding methods.

One of the earliest subjective video quality databases was made available in 2000 when the Video Quality Experts Group (VQEG) finalized the first phase of their FR-TV project [10]. This publicly available database contains 20 standard definition reference videos, each with 16 distortion levels encoded using either MPEG-2 or H.263. In total, the scores of 287 human participants were recorded using the DSCQS methodology. Subsequent databases have provided new content with higher resolutions, updated compression codecs, and different methodologies. For example, the LIVE video database [11], [12] provides 10 reference videos at a resolution of 786 × 432 with 15 different distortions including MPEG-2 and H.264. The more recent IVC [13], [14], IVP [15] databases, and especially the VQEG-HD database [16] provide a large volume of annotated HD interlaced and progressive scan videos. Other video databases of note such as the EPFL [17], [18] and MMSP-SVD [19], [20] are specialized for testing specific spatiotemporal resolutions or network conditions. While many of these databases greatly vary in the resolution and distortions of their sequences, Table I shows how there is very little variation in the lengths of the test sequences they use. The majority of databases use 10-s test sequences, following the recommendation proposed by the ITU, while only one database uses sequences out of the 8–12-s range. For a more detailed account of the most influential subjective VQA databases, comprehensive and insightful accounts can be found in [21] and [22].

TABLE I
Summary of a Selection of Currently Available Subjective Databases With Their Resolution and Test Sequence Duration Statistics. The Majority Adopt the ITU Recommendation of 10 s Video Clips

Database	Source	Duration	Presentation	Resolution
VQEG	20	8s	DS	480i/576i
VQEG-HD	49	10s	SS	1080i/p
LIVE	10	8.68-10s	SS	768×432p
IVP	10	10s	SS	1080p
IVC	24	9-12s	SS	1080i
EPFL	12	10s	SS	Scalable
MMSP-SVD	5	10s	SS+PC	720p
HEVC-CMP1	3	5s	DS	4k
HEVC-CMP2	5	10s	DS	1080p

Double Stimulus (DS), Single Stimulus (SS), Paired Comparison (PC).

As the fidelity of visual display units continues to accelerate, new video coding methods are developed to ensure presentation quality is optimized. The data collected from subjective testing provide essential ground truth with which the performance of these new codecs can be evaluated and compared with that of previous benchmarks. The length of test sequences in subjective databases used to compare new compression formats is not standardized. For example, [2] and [23] (referred to as HEVC-CMP1 and HEVC-CMP2 in Table I) comparing the performance of HEVC with previous methods, employed test sequences of 5 s and 10 s, respectively. Choosing the length of test sequences in such studies is rarely theoretically or empirically motivated, but instead based upon the availability of the desired sequences, which is often limited, especially for ultra-HD content.

B. Variation in Subjective Scores

Fundamentally, the variation existing between subjective video databases falls into two categories: 1) video content and 2) methodological setup. The former is encouraged while the latter is not.

1) Video Content: Winkler [21] argues that the parameterization of video content will facilitate a greater level

of variation in future subjective databases. By extracting values that capture low-level features such as color, spatial edges, and movement, statistics can be calculated that measure content diversity across a selection of videos. This conveniently addresses the problem of content redundancy—a database with a high quantity of videos with similar parameters may be seen as less useful in terms of generalization than a significantly smaller set of videos with a more diverse set of features. By providing such coverage statistics, researchers can not only quantitatively compare subjective databases but also try to maximize content variation when creating new ones. A comprehensive and informative guide to sequence selection for subjective video testing can be found in [24].

2) Methodological Setup: The ITU [3], [4], [25]–[27] outlines several approaches to collecting subjective VQA data. Unfortunately, inconsistencies in these methodologies can lead to undesired variation in opinion scores. Specific attention should be paid to two factors in particular: 1) the stimulus presentation and 2) the rating scale.

The two principal approaches to stimulus presentation in subjective VQA studies differ in the number of videos that are displayed before each recorded response. DS methodologies present videos in pairs, one of which may be a reference or both may be distorted. SS methodologies, on the other hand, present only a single video with no reference to the observer before a response is made. For experiments that use longer clip lengths, SS designs are more suitable as working memory constraints reduce the effectiveness of a DS design. Despite this, researchers often employ SS methodologies for experiments using shorter clips (e.g., the SS continuous quality scale (SSCQS) used for LIVE [11]) due to the significant time savings made compared with collecting data with DS approaches. While the SS methodology is simpler and more efficient than the DS approach, it is also more sensitive to context effects [28].

Context effects occur when previously viewed content influences opinion scores of subsequently viewed footage. In most DS designs, these effects are minimized as relative *difference* opinion scores are recorded between the two presented sequences, as opposed to a single *absolute* opinion score in SS experiments. It should be noted, however, that even DS presentations can be sensitive to context effects in the form of the ordering of the test and the reference video. For example, there is some evidence that observers are more sensitive to degradation in video quality than they are to a similar positive differential in video quality [29]. The impact of such context effects can be reduced, however, by randomly counterbalancing the order of the two sequences. The DSCQS employs such random counterbalancing and has been shown to produce significantly weaker context effects than the DS impairment scale (DSIS) that keeps presentation order fixed [30].

Paired comparison (PC) is a DS presentation methodology that displays two or more videos in parallel or sequentially. In general, each sequence is compared with every other sequence, leading to longer experiment time than most other DS methodologies. Experiments employing the PC methodology facilitate an immediate comparison between multiple sequences. This eliminates concerns over working memory capacity when using longer sequences and results have been reported that claim to show the superiority of PC over SSCQS [20]. However, SSCQS still requires fewer trials than PC as usually all pairs of videos need evaluation in PC, whereas SSCQS multiple test trials can be compared with a single reference trial. Despite this, procedures exist to reduce the number of PC comparisons [31].

Subjective assessment methodology for video quality (SAMVIQ) [26] is a testing methodology that differs from those previously described here in the level of control and freedom afforded the observer. Viewers are presented with a series of scenes, providing viewers with an interface with a single video window and playback controls allowing the observer to view an explicit reference sequence and several different distorted versions as many times as they choose before providing an onscreen response to each. A study comparing an SS methodology [absolute category rating (ACR)] and SAMVIQ indicated that SAMVIQ provides a greater level of accuracy than ACR if data are collected with the same number of observers [32]. The benefits of increased accuracy using the SAMVIQ methodology are tempered by the additional time costs associated with the open-ended interactivity of the paradigm, over single and DS approaches and the artificial nature of the viewing experience [33].

Different rating scales that collect opinion scores in VQA studies may also be a source of undesired variability. Scales can vary with respect to being continuous or discrete, in the number of rating points and in the labels that accompany them. Continuous scales (such as those used in DSCQS and SSCQS) ask observers to use an onscreen slider or line bisection to record their response, whereas discrete scales (such as the DSIS or ACR) require participants to choose one of a finite number of ordinal quality levels. Continuous scales have the advantage of providing a more accurate evaluation response but may also be more sensitive to spatial biases as discrete scales have been reported to provide more stable data [34].

Invariably, both discrete and continuous scales are accompanied with descriptive words that can potentially introduce confusion and unwanted variation in opinion scores. For example, the DSCQS asks observers to rate videos on a scale of 0–100 where 0 is *bad* and 100 is *excellent*, while DSIS asks observers to rate distortions on a five-level scale that range from *very annoying* to *imperceptible*. The nonlinear mapping between discrete and continuous scores is a further reason that makes a comparison of the two scales problematic. In discrete scales, the five standard ITU labels in English are: 1) *excellent*; 2) *good*; 3) *fair*; 4) *poor*; and 5) *bad*. It may be assumed that when placed upon a continuous scale, these words are positioned with equal intervals. This assumption has been questioned, however, by studies that indicate that the semantic intervals between the ITU labels vary dependingly on the language they are expressed in [35] and [36]. This means that discrete scores cannot be easily converted and compared with scores from continuous-scale experiments. While different scales may be necessary in specific contexts, they make comparisons between databases difficult, and furthermore, varying the language of the labels may introduce additional noise.

Fig. 1. Individual frames taken from each of the four reference sequences used in this paper.

Despite these theoretical reservations, it should also be noted that recent research comparing the effects of four different scales found strong correlations between continuous and discrete variants, specifically in the context of the SS methodology [37].

C. Previous Research on Clip Duration in VQA

Surprisingly, the majority of previous research on test sequence lengths in subjective VQA has explored the impact of varying durations above 10 s rather than below it. Reference [5] has suggested that longer sequence lengths reduce human sensitivity to compression artifacts [5]. An SS design was used to compare the opinion scores of a group of three shorter length sequences (10, 15, and 30 s) with those of a group of three longer length sequences (60, 120, and 240 s) and reported significantly higher MOSs for the group of longer sequences but found no significant differences within each of the groups. The researchers suggest when watching longer video sequences, observers focus more upon the content of the sequences, and in doing so, become less critical of the presentation. This interpretation is supported by the effect being strongest in the highest quality (lowest compression) content. As no significant effects were found between the groups of 10, 15, and 30 s, the researchers supported the current ITU recommendation of 10 s; however, it is not clear whether observers continue to become more critical if durations are reduced below 10 s.

One reason to believe that they might do is that recency effects have been found indicating observers converge upon a quality assessment after viewing around 6 s of a video clip. Pinson and Wolf [28] correlated opinion scores collected at the *end* of a viewing with continuous scores collected during a viewing. They found that the correlation coefficient values moved above 0.9 after 6 s of viewing the video. Aldridge et al. [38] found that while viewing 30-s clips, observers weighted their evaluation of the final 10 s 10% more than footage seen in the first 10 s [38].

Shorter sequence durations are also more likely to produce consistent ratings between observers than longer ones. This is because there is a greater level of consistency in where people look immediately after the onset of a scene, with a rapid increase in variation after 2–3 s [6], [7]. If there is variance in where people are attending, there is likely to be associated variance in their opinion scores.

It has been argued elsewhere [5] that sequence lengths in subjective VQA studies should be longer than 10 s as the natural conditions under which people watch videos are, invariably, much longer. A counterargument to this point is that it is not the length of the natural viewing period that needs to be emulated in test conditions, but instead the length of individual shots. While there remain notable types of video content that consistently use longer shot lengths (e.g., video conferencing or sporting event coverage), modern cinema typically employs shot lengths significantly shorter than 10 s. Reference [8], that examined the average shot length of Hollywood movies over the last 75 years, discovered a strong trend that saw average shot lengths fall from 10 s in the 1930s to below 4 s in the 2000s. Therefore, for subjective VQA studies to emulate the experience of contemporary cinema goers, shorter sequence lengths should be considered.

III. Methodology

This section contains a detailed specification of a subjective experiment, designed to explore the impact of reducing video presentation times below 10 s, upon the criticality of observer rating behavior.

A. Participants

Twenty-four postgraduate students (11 women and 13 men) at the University of Bristol were paid to participate. The average age of participants was 23.7 years and all reported having normal or corrected-to-normal visual acuity. All also had normal color vision, which was verified by the use of Ishihara charts.

B. Reference Sequences

Four HD, uncompressed reference sequences were selected from the VQEG HD database [16]. Each of the chosen videos was originally 1920×1080 pixels, transformed to YUV 4:2:0 format, progressive scan and played at 25 frames/s. All original reference sequences were 10-s long and contained no shot transitions or audio components. Sequences were chosen that maintained a high level of temporal consistency throughout the entire 10 s to ensure that content did not significantly vary when trimmed to shorter lengths. A description of the visual characteristics of the source videos is provided below and example frames taken from each can be observed in Fig. 1. The types of movement featured in each sequence are described in Table II.

1) *Abbey:* Sequence shot from an aerial perspective slowly moving toward a large house surrounded by grass and trees.

TABLE II
DESCRIPTION OF TYPES OF MOVEMENT FEATURED IN THE FOUR REFERENCE SEQUENCES

Sequence	Camera	Structured Movement	Dynamic texture
Abbey	Pan + Zoom	NA	Tree leaves
Bottles	Zoom	NA	NA
Feathers	Static	Rotating cage	Moving feathers
Waves	Static	NA	Rippling water

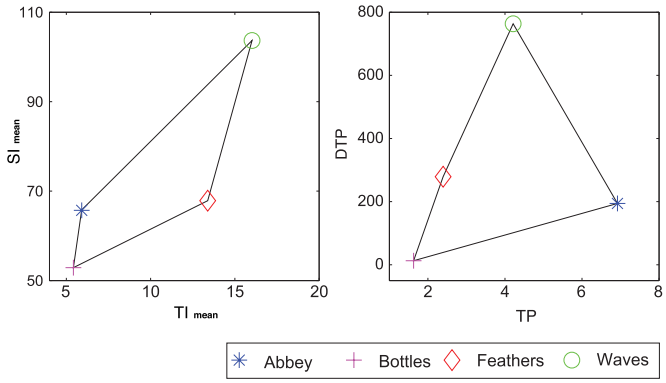

Fig. 2. Plots of the average feature values of the four reference sequences over the full 10 s.

2) *Bottles:* Camera zooms out slowly then maintains a static shot of an arrangement of Bottles and cans.
3) *Feathers:* A static camera views a revolving metallic cage decorated with colorful feathers.
4) *Waves:* Sequence shot from a static camera depicting rippling water beside a glistening rockface.

Four low-level feature descriptors were also computed [21], [39] for each reference video to quantify the content of the database. These features are described in detail in the Appendix. Mean spatial information (SI_{mean}) is an estimation of edge density [21]. Mean temporal information (TI_{mean}) is calculated as the absolute difference in intensity of every frame in a sequence. The texture parameter (TP) [39] and dynamic TP (DTP) [39] describe static and dynamic texture properties, respectively. The average coverage of the features calculated using the current four sequences can be seen in Fig. 2, while the time-variant plots are displayed in Fig. 4.

C. Test Sequences

For the test material, five versions of each reference sequence were generated, their durations 10, 7, 5, 3, and 1.5 s, respectively. Shorter sequences were created by trimming the necessary amount of time from the end of the 10-s sequence, unless the video featured a significant shift in camera dynamics. In this case, the shorter sequences were produced by symmetrically trimming either side of the event. One such event was identified in the *Bottles* video whereby the camera switched from a zoom-out to static. Fig. 4 plots the time-varying feature values for the four reference sequences with an indication of which frames were included in the truncated versions.

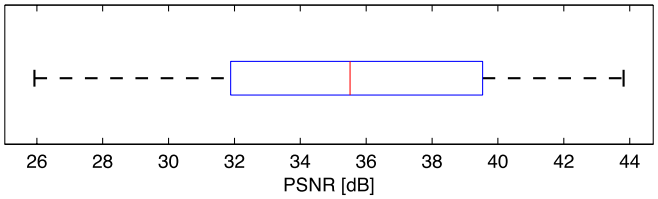

Fig. 3. Boxplot illustrating the distribution of PSNR values calculated from the group of twenty 10-s test sequences. The red line represents the median, the vertical edges of the box are the first and third quartiles, and the whiskers are the range.

Each video sequence was also distorted using HEVC compression (codec version HM 14.0) with four different quantization parameters (QPs) of 27, 32, 37, and 42. Maximum coding units were set to 64 × 64 pixels, maximum partition depth set to 4, and the group of pictures size was 1. One additional distortion was generated using a Gaussian blurring filter of 7 × 7 pixels with a standard deviation of 4 pixels. The distribution of peak signal to noise ratio (PSNR) values for the twenty 10 s test sequences can be seen in Fig. 3. The final data set consisted of 20 reference videos and 100 test videos.

D. Environmental Setup

All videos were displayed at 25 frames/s on a Panasonic TH-50BT300ER HD plasma screen, screen with a native resolution of 1920 × 1080 pixels and measuring 1105 × 622 mm and a 5 000 000:1 contrast ratio. The screen was connected to a Windows PC running MATLAB and Psychtoolbox 3.0. Participants sat in a chair 186.6 cm from the screen (three times the height of the screen) in a darkened room.

E. Assessment

Participants provided three responses after viewing each video pair using a physical questionnaire sheet and a pencil. First, participants indicated which of the two videos was of inferior quality. Then, second and third, participants provided a quality score for the first and second videos. Scores were expressed by line bisection of a quality scale labeled at equal intervals with *excellent*, *good*, *fair*, *poor*, and *bad*.

F. Procedure

The DSCQS procedure was used to collect the data. Each trial consisted of the participant viewing Video A followed by a gray screen for 3000 ms then viewing Video B. Video B was followed by a gray screen depicting a 3000-ms countdown. Participants were then asked to produce a subjective quality evaluation of both Videos A and B using the assessment method described in the above section. Participants were given unlimited time to make their choice before beginning the next trial.

For each pair of videos, one was a reference, while the other was a distorted version of the reference. The order of the reference and test sequences was randomly counter balanced. Trials were grouped according to sequence length, producing five blocks of trials. Within each of these, participants viewed

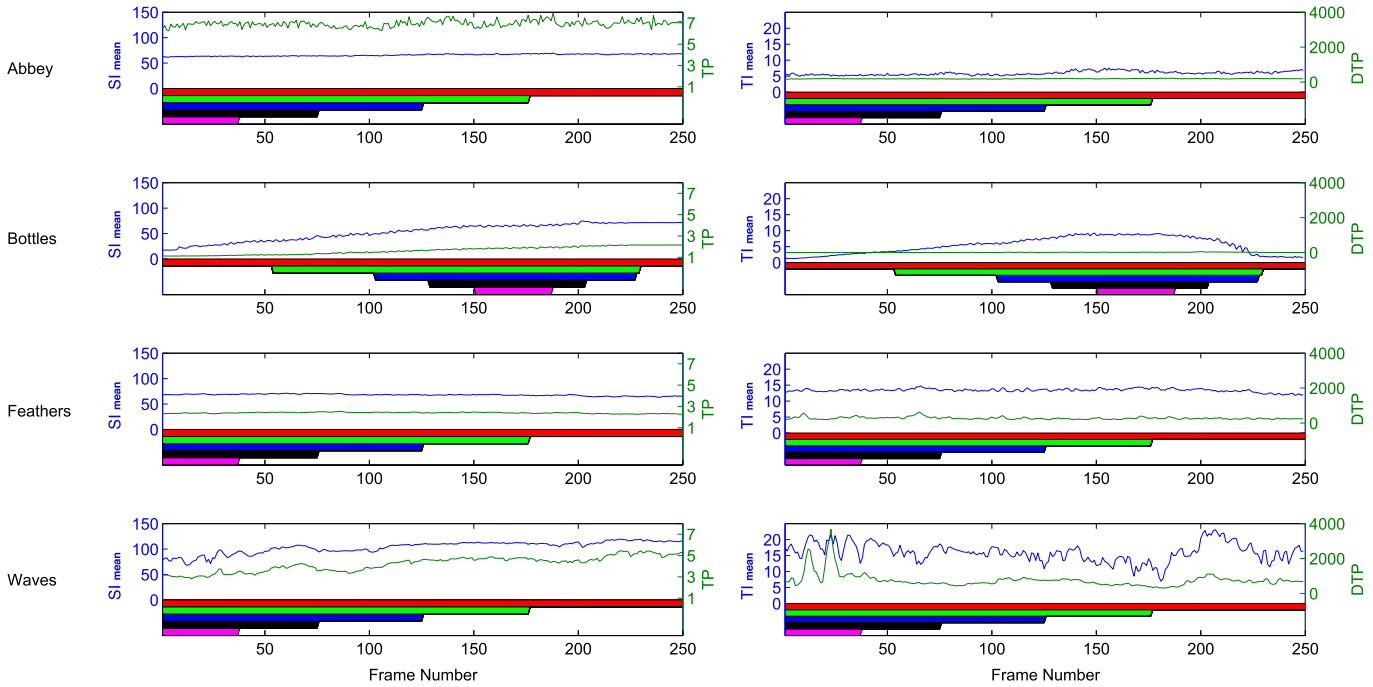

Fig. 4. Time-varying feature plots for each of the four reference sequences used in this paper. The colored bars at the base of plots indicate the frames used in truncated sequences: red (10 s), green (7 s), blue (5 s), black (3 s), and magenta (1.5 s). Care was taken to ensure consistency of feature content across different duration sequences.

sequences of only a single sequence duration. The order of the blocks and the order of trials within each block were randomly permutated for each participant to prevent ordering effects.

In total, each participant viewed five blocks, containing 20 trials each with a complete session lasting no more than 30 min.

Prior to testing, each participant was given instructions and took part in a brief training session providing opinion scores for two video pairs, each of 10 s duration. The videos used in the training session were not used in the subsequent test session.

After completing the five blocks of trials, observers were asked the question: "In which of the five blocks did you feel comfortable making quality assessments?.

G. Analysis

Difference scores were calculated for each trial and each participant by subtracting the quality score (measured in centimeters and scaled up to a value between 0 and 100) of the test sequence from that of the paired reference sequence. Difference MOSs (DMOSs) were calculated for every participant by taking their average difference score for trials within each duration block (collapsing over different distortions and sequences).

Participant data were checked for outliers according to the protocol outlined by the ITU [3]. Here, values that fall two standard deviations above or below the mean value are considered outliers. A participant from a normally distributed population (defined as distribution with a Kurtosis value between 2 and 4) is rejected if two conditions are met. The first condition is met if over 5% of a participant's scores are outliers. The second condition is met if the ratio of the absolute difference between the number of high and low outliers to the total number of outliers of a participant falls below 30%. This analysis indicated that one participant should be removed from the data set (6% for criteria 1 and 0% for criteria 2).

IV. Results and Discussion

This section is divided into four parts, each of which addresses a separate research question posed in Section I. Three main analysis techniques are employed: 1) analysis of variance (ANOVA); 2) pairwise comparisons, adjusted for multiple comparisons using Tukey's least significant difference; and 3) Pearson's correlation coefficient [40]. For a comprehensive guide to statistical analysis of subjective testing data, the reader is referred to ITU-T Study Group 12 [27]. Section IV-A uses a one-way ANOVA to address the main question of whether variance in DMOS exists between different sequence lengths. To identify whether compression level or reference video had a significant influence upon a potential duration effect, a two-way ANOVA is used in Sections IV-B and IV-C, respectively. The significance level of the interaction effect in a two-way ANOVA provides an indication of whether the main effect of the first factor is significantly different under distinct conditions introduced by the second factor. If an ANOVA is significant overall, pairwise comparisons are used to identify which pairs of durations yield significantly different DMOSs. The significance of a Pearson's correlation coefficient is used to identify whether there is a linear relationship between sequence duration and DMOS.

In Section IV-A, analyses were performed upon DMOSs from all videos and all compression levels.

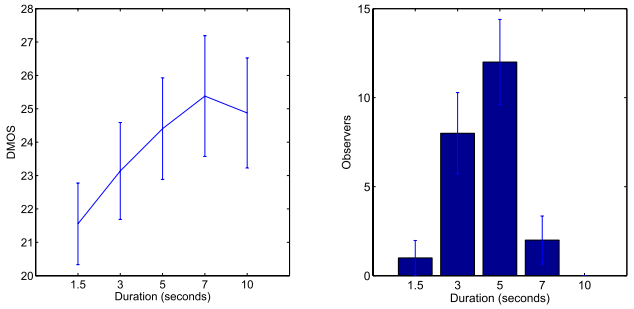

Fig. 5. (a) When averaging DMOS values over all four reference sequences and five compression strengths, a smooth function emerges, indicating longer sequences produce higher levels of accuracy. The error bars represent the standard error of the mean. (b) Histogram of the observer votes for the shortest duration sequence they felt confident assessing. The error bars represent the standard error of the mean.

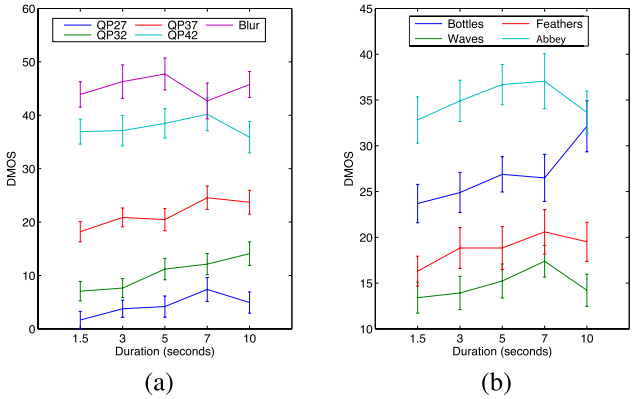

Fig. 6. (a) Separating DMOS with respect to compression strength highlights a distinction between trials using the two most aggressively compressed sequences and the remaining three. The error bars represent the standard error of the mean. (b) Separating DMOS with respect to reference sequence highlights how the main effect of an increase between 1.5-s sequences and 7-s sequences remains consistent for all four videos in the database. The error bars represent the standard error of the mean.

In Sections IV-B and IV-C, analyses were performed after separating DMOS values according to compression level and reference video, respectively.

A. Sequence Duration Affects DMOSs

The average DMOS over all conditions and participants was 24, indicating that scores were not uniformly distributed over the 100-point scale. Fig. 5(a) combines DMOS from every sequence and all distortion types and illustrates how the accuracy of human observers steadily increased when viewing sequences from the 1.5-s block to sequences from the 7-s block, before decreasing slightly during the 10-s block. However, it must also be noted that the differences are very small. A one-way repeated measures ANOVA performed upon participant DMOS produced a significant model, $F(488) = 2.78$ and $p = .032$, indicating that significant variation existed between the groups. After adjusting for multiple comparisons, significant differences were found between the block of 1.5-s sequences and those of 5-s ($p = .036$), 7-s ($p = .024$), and 10-s sequences ($p = .036$). Pairwise comparisons between DMOS values in the 5-, 7-, and 10-s blocks were not significant (all $p > .48$). Correlational analysis confirmed that an overall linear relationship between duration and DMOS was not significant ($r(21) = 0.16$ and $p = .088$). These results indicate that longer durations do increase observers' accuracy when identifying compression artifacts; however, the effect is a small one and 10 s did not produce the best performance.

B. Influence of Compression Strength

The principle point to highlight about the plot in Fig. 6(a) is that as expected, different compression levels produced different strengths of DMOS values. Higher compression levels increase the amount of distortion between the reference and test videos and this is reflected in the DMOS values. A two-way repeated measures ANOVA (factor 1: sequence duration and factor 2: video compression) confirmed this by identifying compression as a significant factor affecting the variance of DMOS, $F(2.23, 49.21) = 140.35$, and $p < .001$ (degrees of freedom adjusted using the Greenhouse–Geisser correction due to violation of sphericity assumption). The model yielded a marginally significant interaction between compression and duration, $F(16\,352) = 1.62$, and $p = .06$, suggesting that the duration effect identified in the previous section varied as a function of compression level. Inspection of Fig. 6(a) indicates that the three least aggressive compression levels (QP27, QP32, and QP37) produced the most consistent trend. This observation was confirmed by independent analysis of the DMOS at each QP value (the results are shown in Table III). Videos compressed at QP32 and QP37 produced significant overall ANOVA models and significant correlations between sequence duration and DMOS. Furthermore, significant or marginally significant pairwise comparisons appeared only in the QP27 model (1.5 versus 7 s and $p = .04$), the QP32 model (1.5 versus 5 s and $p = .064$, 1.5 versus 7 s and $p = .019$, 1.5 versus 10 s and $p = .018$) and the QP37 model (1.5 versus 7 s and $p = .02$, 1.5 versus 10 s and $p = .05$). Data from trials using videos with the two most aggressive compressions (QP42 and Blur) produced no significant effects with respect to duration.

TABLE III
ANOVA AND CORRELATION STATISTICS SEPARATED BY COMPRESSION LEVEL. STATISTICS SIGNIFICANT TO THE $p < .05$ LEVEL IN BOLD WITH ASTERISK

QP	$F(4,88)$	p	$r(21)$	p
27	1.35	0.26	0.14	0.13
32	**4.35***	**0.03***	**0.27***	**0.003***
37	**2.59***	**0.04***	**0.20***	**0.03***
42	0.96	0.43	0.002	0.99
Blur	1.47	0.22	0.001	0.99

Both the level of compression and the length of the test sequence affect the amount of information available to the observer when completing the task. The results here indicate that the benefits gained by longer durations are tempered if the artifact detection task is either too easy or too hard. In the case of the task being too easy (the two highest levels of distortion), there is already enough information provided in each of the

TABLE IV
ANOVA AND CORRELATION STATISTICS SEPARATED AT REFERENCE SEQUENCE LEVEL. STATISTICS SIGNIFICANT TO THE $p < .05$ LEVEL IN BOLD WITH ASTERISK. † DEGREES OF FREEDOM ADJUSTED TO $F(2.81, 61.73)$ USING THE GREENHOUSE–GEISSER CORRECTION DUE TO VIOLATION OF SPHERICITY ASSUMPTION

Sequence	$F(4,88)$	p	$r(21)$	p
Abbey	0.97†	0.41	0.03	0.76
Bottles	**4.33***	**0.003***	**0.24***	**0.01***
Feathers	0.93	0.45	0.1	0.27
Waves	1.23	0.31	0.07	0.46

sequence lengths to make an accurate decision. By contrast, in the case of the task being too difficult (the lowest level of distortion), the scarcity of signal separating the reference from the test sequence reduces the impact of a longer viewing time.

This suggests that when identifying compression artifacts, the benefits afforded to viewing longer sequence durations are not as valuable if the task is very difficult or very easy.

C. Influence of Video Content

Interestingly, the DMOS for each of the four reference videos displayed in Fig. 6(b) was significantly different. A two-way repeated measures ANOVA (factor 1: sequence duration and factor 2: reference sequence) indicated source video to be a highly significant factor contributing to variance in DMOS, $F(366) = 57.92$, and $p < .001$. All four videos produced similar increasing trends from the 1.5-s block to sequences from the 7-s block after which, for the 10-s block, the DMOS declined in all but the *Bottles* sequence, which conversely, produced a steep increase. Despite the anomalous increase observed for the *Bottles* sequence, the interaction effect was not significant, indicating that, overall, the effect did not change in different test sequences.

The results of independent analyses performed upon the data from each video sequence are displayed in Table IV. While data from the *Bottles* video produced highly significant ANOVA and correlation models, all other videos produced no significant effects. Examining the pairwise comparisons, one saw this trend continuing with data from the *Bottles* sequence displaying a highly significant increase between 1.5 and 10 s ($p = .003$), whereas all other videos failed to yield any significant differences. So, what was different about the *Bottles* video compared with the others?

The current data set is not diverse enough to draw strong conclusions about why DMOS for the *Bottles* video continued to increase in the 10 s block, while this was not the case for the other three videos. However, an examination of the feature plots in Fig. 4 suggests the 10-s *Bottles* sequence has specific properties that may explain its irregular results. Critically, the point at which the TI_{mean} drops off to zero corresponds to the point at which the 7 s sequence ends, while the 10-s sequence continues for a further 3 s. Each of the other three 10-s sequences features significant movement throughout their durations. If observers are weighting their observations favorably toward the end of the sequences (as suggested in [28] and [38]), an interpretation of the present data is that the static camera at the end of the *Bottles* video provided a critical advantage to observers viewing the full 10-s *Bottles* video. The presence of static text toward the end of the 10-s *Bottles* video further supports this interpretation. The criticality task is likely to become easier for observers when viewing static text as minor distortions to this kind of content are more salient than those to less predictable or structured stimuli. If our speculations here are true, the increase in DMOS between the 7 and 10-s *Bottles* sequence was driven not by the increase in duration but instead by the decrease in movement.

Three of the four test sequences produced the same pattern seen in the overall duration effect, while the fourth sequence was very similar. Therefore, the results here indicate that video content has a minimal impact on the way duration affects DMOS values.

D. Observer Assessment Confidence

After testing was complete, observers were asked in which of the blocks they felt confident completing the task. Fig. 5(b) shows how the majority of observers (12) identified 5 s to be the shortest duration sequence that they felt confident assessing. Eight observers identified 3-s sequences, two observers chose 7-s sequences and one observer selected 1.5-s sequences. Interestingly, not one of the participants chose 10-s sequences, recommended by the ITU and used by the majority of current subjective VQA studies. A chi-squared test of independence confirmed that these frequencies are significantly different, $\chi^2(4) = 23.3$ and $p < .001$. Observers, therefore, felt just as confident in their quality judgments while watching shorter sequence lengths as longer ones, with 3- and 5-s sequences, accounting for 87% of the votes.

V. CONCLUSION

Here, we have reported the results of a subjective VQA study that explored the impact on rating behavior of reducing test sequence durations below the standard 10 s, recommended by the ITU. Our four significant findings are the following.

1) There is a small but significant increase in accuracy if sequences are increased from 1.5 to 5, 7, or 10 s.
2) This effect becomes stronger if the difference in distortion between the reference and test video is reduced.
3) The main effect remains consistent between different but temporally consistent source videos.
4) Observers feel just as confident assessing the quality of videos that are 5 s as ones that are 10 s.

The practical implications of these findings are significant. Our results indicate that critical observations of video quality do not significantly change if 10-s sequences are exchanged for 7-, or indeed, 5-s sequences. However, our recommendation to half the standard sequence length from 10 to 5 s is qualified by the methodology used for data collection and the temporal consistency of the content. The findings presented here specifically relate to studies that use DS presentations and the continuous quality scale assessment technique. For experiments that deviate from this particular design, our recommendations are not necessarily applicable. A final qualifying note to the reader is that the current results are based on the analysis of a diverse set of content, but the number of sequences was

constrained to ensure a manageable testing time. Three of the four videos produced very similar results; however, the one featuring the most spatiotemporal variation over time (*Bottles*) slightly deviated from the common pattern. Our strongest recommendation can, therefore, only be applied to temporally consistent content. However, future research may focus on whether the findings reported here do indeed translate to other methodologies and more varied content.

Collecting ground truth data is essential for the comparison of video coding methods and for validation of objective VQA models, but it is expensive, in both time and labor. The results presented here provide convincing evidence that by reducing sequence lengths from 10 to 5 s, video processing times can be halved and test time can be cut by 40% in DS methodologies (assuming an additional 5 s for voting), without any significant impact on the quality of the resulting data. If voting times can be reduced, the potential savings increase. Furthermore, the time to compress test sequences will also be reduced proportionally. We believe that such a shift in procedure will facilitate a significant boost in the generation of ground truth data produced by the associated research community.

APPENDIX
FEATURE CALCULATIONS

Mean spatial information ($\mathrm{SI}_{\mathrm{mean}}$) is an estimation of edge density. It is calculated for each frame by convolving the luminance profile with Sobel filters [41]. Images filtered with the horizontal, $I_h(x, y)$, and vertical, $I_v(x, y)$, Sobel kernels are then combined to produce an image representing edge magnitude at each pixel. The final $\mathrm{SI}_{\mathrm{mean}}$ frame descriptor is calculated as the root mean square of edge magnitude at each pixel

$$\mathrm{SI}_{\mathrm{mean}} = \sum_{x,y} \frac{\sqrt{I_h(x, y)^2 + I_v(x, y)^2}}{P} \tag{1}$$

where P is the total number of pixels in the frame.[2]

Mean temporal information ($\mathrm{TI}_{\mathrm{mean}}$) is an estimation of variation in luminance between neighboring frames of a sequence. It is calculated as the root mean square difference in luminance between the current frame $I_t(x, y)$ and the previous frame $I_{t-1}(x, y)$

$$\mathrm{TI}_{\mathrm{mean}} = \sum_{x,y} \frac{\sqrt{(I_t(x, y) - I_{t-1}(x, y))^2}}{P}. \tag{2}$$

TP is an estimation of static texture and is calculated under the assumption that texture resides in regions dominated by high-spatial-frequency components. Each frame is decomposed into six high-frequency subbands, using the dual-tree wavelet transformation [42]. Subband coefficients, $B_{1:6}(x, y)$ are then summed to produce a single TP map

$$\mathrm{M}_{tp}(x, y) = \sum_{i=1}^{6} B_i(x, y). \tag{3}$$

The final TP value for a single frame is obtained by calculating the mean over all pixels

$$\mathrm{TP} = \sum_{x,y} \frac{M_{tp}(x, y)}{P}. \tag{4}$$

DTP is an estimation of complex and irregular motion between the current and two reference frames [39]. DTP is calculated as the product of two features, DTP_1 and DTP_2

$$\mathrm{DTP} = \mathrm{DTP}_1 \cdot \mathrm{DTP}_2. \tag{5}$$

Motion estimation is applied between a current and reference frame based on a translational model that uses sum of squared differences as the distortion method, 8×8 blocks, and a full search strategy (64 pixel range). The discrete approximation for the second derivative [$\mathrm{SD}(x, y)$, calculated based on two subvectors $\mathbf{SDX}(x, y)$ and $\mathbf{SDY}(x, y)$] of a motion vector $\mathrm{MV}(x, y)$, is formally expressed in (6) and (7). It is noted here that (x, y) in these equations refer to block coordinates as opposed to pixel coordinates, referenced in previous equations[3]

$$\mathbf{SDX}(x, y) = \mathbf{MV}(x - 1, y) + \mathbf{MV}(x + 1, y) - 2\mathbf{MV}(x, y) \tag{6}$$

$$\mathbf{SDY}(x, y) = \mathbf{MV}(x, y - 1) + \mathbf{MV}(x, y + 1) - 2\mathbf{MV}(x, y) \tag{7}$$

where $\mathbf{MV}(x, y) = (\mathrm{MV}_h(x, y), \mathrm{MV}_v(x, y))$ is the motion vector of an 8×8 block. Here, motion vectors are calculated between the current and two adjacent reference frames. For each reference frame p, the $\mathrm{SDX}(x, y)$ and $\mathrm{SDY}(x, y)$ are combined according to

$$\mathrm{SD}_p(x, y) = ||\mathbf{SDX}_p(x, y)||_2 + ||\mathbf{SDY}_p(x, y)||_2. \tag{8}$$

The descriptors for each of the two reference frames are then combined, weighted by their distance from the current frame

$$\mathrm{SD}(x, y) = \sum_{p=\pm 1} \frac{\frac{1}{|p|} \cdot \mathrm{SD}_p(x, y)}{2}. \tag{9}$$

The final DTP_1 feature is then calculated by taking the mean value over all blocks

$$\mathrm{DTP}_1 = \sum_{x,y} \frac{\mathrm{SD}(x, y)}{N} \tag{10}$$

where x and y are the block coordinates and N is the total number of blocks in a single frame.

For DTP_2, the mean squared error in luminance values between the current frame $I_c(x, y)$ and the motion-compensated frame $I_p(x, y)$ is calculated before averaging over all pixels in the frame

$$\mathrm{MSE}_p = \sum_{x,y} \frac{\sqrt{((I_c(x, y) - I_p(x, y))^2}}{P} \tag{11}$$

[2]Unless otherwise stated, we employ capital letters to represent matrices, for example, I_h is the matrix representing a single frame after convolution with the horizontal Sobel kernel. The capital letters with coordinates represent matrix elements, for example, $I_h(x, y)$.

[3]We use bold capital letters to represent the matrices with vector elements, for example, **MV**. We use bold capital letters with coordinates to represent the elements of these matrices, for example, $\mathbf{MV}(x, y)$.

where MSE_p is the mean squared error in luminance between the current frame and reference frame p. For the final DTP_2 feature value, mean squared errors calculated from each reference frame are combined, weighted by the distance from the current frame

$$\mathrm{DTP}_2 = \sum_{p=\pm 1} \frac{\frac{1}{|p|} \cdot \mathrm{MSE}_p}{2}. \tag{12}$$

REFERENCES

[1] D. R. Bull, *Communicating Pictures: A Course in Image and Video Coding*, 1st ed. Oxford, U.K.: Academic, 2014.

[2] J. Ohm, G. J. Sullivan, H. Schwarz, T. K. Tan, and T. Wiegand, "Comparison of the coding efficiency of video coding standards—Including High Efficiency Video Coding (HEVC)," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 12, pp. 1669–1684, Dec. 2012.

[3] *Methodology for the Subjective Assessment of the Quality of Television Pictures*, document ITU-R BT.500-11, Geneva, Switzerland, 2002.

[4] *Methodology for the Subjective Assessment of the Quality of Television Pictures*, document ITU-R BT.500-13, Geneva, Switzerland, Jan. 2012.

[5] P. Frohlich, S. Egger, R. Schatz, M. Muhlegger, K. Masuch, and B. Gardlo, "QoE in 10 seconds: Are short video clip lengths sufficient for quality of experience assessment?" in *Proc. 4th Int. Workshop Quality Multimedia Exper. (QoMEX)*, Jul. 2012, pp. 242–247.

[6] B. W. Tatler, R. J. Baddeley, and I. D. Gilchrist, "Visual correlates of fixation selection: Effects of scale and time," *Vis. Res.*, vol. 45, no. 5, pp. 643–659, Mar. 2005.

[7] F. J. M. Moss, R. J. Baddeley, and N. Canagarajah, "Eye movements to natural images as a function of sex and personality," *PLoS One*, vol. 7, no. 11, p. e47870, Nov. 2012.

[8] J. E. Cutting, K. L. Brunick, J. E. DeLong, C. Iricinschi, and A. Candan, "Quicker, faster, darker: Changes in Hollywood film over 75 years," *i-Perception*, vol. 2, no. 6, pp. 569–576, 2011.

[9] G. J. Sullivan, J.-R. Ohm, W.-J. Han, and T. Wiegand, "Overview of the High Efficiency Video Coding (HEVC) standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 12, pp. 1649–1668, Dec. 2012.

[10] *Final Report From the Video Quality Experts Group on the Validation of Objective Quality Metrics for Video Quality Assessment*, Jun. 2000. [Online]. Available: www.vgeq.org

[11] K. Seshadrinathan, R. Soundararajan, A. C. Bovik, and L. K. Cormack, "Study of subjective and objective quality assessment of video," *IEEE Trans. Image Process.*, vol. 19, no. 6, pp. 1427–1441, Jan. 2010.

[12] K. Seshadrinathan, A. C. Bovik, and L. K. Cormack. (2010). *LIVE Video Quality Database*. [Online]. Available: http://live.ece.utexas.edu/research/quality/live_video.html

[13] S. Péchard, R. Pépion, and P. Le Callet. *IRCCyN/IVC Videos 1080i Database*. [Online]. Available: http://www.irccyn.ec-nantes.fr/~lecallet/platforms.htm, accessed Oct. 16, 2015.

[14] S. Péchard, R. Pépion, and P. Le Callet, "Suitable methodology in subjective video quality assessment: A resolution dependent paradigm," in *Proc. Int. Workshop Image Media Quality Appl.*, Sep. 2008, p. 6.

[15] F. Zhang, S. Li, L. Ma, Y. C. Wong, and K. N. Ngan. *IVP Subjective Quality Video Database*. [Online]. Available: http://ivp.ee.cuhk.edu.hk/research/database/subjective/, accessed Oct. 16, 2015.

[16] *Report on the Validation of Video Quality Models for High Definition Video Content*, Jun. 2010. [Online]. Available: www.vgeq.org

[17] F. De Simone, M. Naccari, M. Tagliasacchi, F. Dufaux, S. Tubaro, and T. Ebrahimi, "Subjective assessment of H.264/AVC video sequences transmitted over a noisy channel," in *Proc. Int. Workshop Quality Multimedia Exper.*, Jul. 2009, pp. 204–209.

[18] F. De Simone, M. Tagliasacchi, M. Naccari, S. Tubaro, and T. Ebrahimi, "A H.264/AVC video database for the evaluation of quality metrics," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Dallas, TX, USA, Mar. 2010, pp. 2430–2433.

[19] J.-S. Lee *et al*. (2010). *MMSP Scalable Video Database*. [Online]. Available: http://mmspg.epfl.ch/svd

[20] J.-S. Lee, F. De Simone, and T. Ebrahimi, "Subjective quality evaluation via paired comparison: Application to scalable video coding," *IEEE Trans. Multimedia*, vol. 13, no. 5, pp. 882–893, Oct. 2011.

[21] S. Winkler, "Analysis of public image and video databases for quality assessment," *IEEE J. Sel. Topics Signal Process.*, vol. 6, no. 6, pp. 616–625, Oct. 2012.

[22] K. Fliegel and C. Timmerer, *WG4 Databases White Paper v1.5: QUALINET Multimedia Database Enabling QoE Evaluations Benchmarking*, document Qi0306 Prague, Czech Republic, Mar. 2013.

[23] P. Hanhart, M. Rerabek, F. De Simone, and T. Ebrahimi, "Subjective quality evaluation of the upcoming HEVC video compression standard," *Proc. SPIE, Appl. Digit. Image Process. XXXV*, vol. 8499, pp. 84990V-1–84990V-13, Oct. 2012.

[24] M. H. Pinson, M. Barkowsky, and P. Le Callet, "Selecting scenes for 2D and 3D subjective video quality tests," *EURASIP J. Image Video Process.*, vol. 2013, no. 1, p. 50, Aug. 2013.

[25] *Subjective Video Quality Assessment Methods for Multimedia Applications*, document ITU-T P.910, Geneva, Switzerland, Sep. 1999.

[26] *Methodology for the Subjective Assessment of Video Quality in Multimedia Applications*, document ITU-R BT.1788, Geneva, Switzerland, Feb. 2007.

[27] *Methods, Metrics and Procedures for Statistical Evaluation, Qualification and Comparison of Objective Quality Prediction Models*, document ITU-T P.1401, Geneva, Switzerland, Jul. 2012.

[28] M. H. Pinson and S. Wolf, "Comparing subjective video quality testing methodologies," *Proc. SPIE, Vis. Commun. Image Process.*, vol. 5150, pp. 573–582, Jun. 2003.

[29] R. Hamberg and H. de Ridder, "Time-varying image quality: Modeling the relation between instantaneous and overall quality," *SMPTE J.*, vol. 108, no. 11, pp. 802–811, Nov. 1999.

[30] P. Corriveau, C. Gojmerac, B. Hughes, and L. Stelmach, "All subjective scales are not created equal: The effects of context on different scales," *Signal Process.*, vol. 77, no. 1, pp. 1–9, Aug. 1999.

[31] J. Li, M. Barkowsky, and P. Le Callet, "Boosting paired comparison methodology in measuring visual discomfort of 3DTV: Performances of three different designs," *Proc. SPIE, Stereosc. Displays Appl. XXIV*, vol. 8648, pp. 86481V-1–86481V-13, Mar. 2013.

[32] D. M. Rouse, R. Pépion, P. Le Callet, and S. S. Hemami, "Tradeoffs in subjective testing methods for image and video quality assessment," *Proc. SPIE, Human Vis. Electron. Imag. XV*, vol. 7527, pp. 75270F-1–75270F-11, Feb. 2010.

[33] Q. Huynh-Thu, M. Brotherton, D. Hands, K. Brunnström, and M. Ghanbari, "Examination of the SAMVIQ methodology for the subjective assessment of multimedia quality," in *Proc. 3rd Int. Workshop Video Process. Consum. Electron.*, Scottsdale, AZ, USA, Jan. 2007.

[34] E. Svensson, "Comparison of the quality of assessments using continuous and discrete ordinal rating scales," *Biometrical J.*, vol. 42, no. 4, pp. 417–434, 2000.

[35] B. L. Jones and P. R. McManus, "Graphic scaling of qualitative terms," *SMPTE J.*, vol. 95, no. 11, pp. 1166–1171, Nov. 1986.

[36] M. T. Virtanen, N. Gleiss, and M. Goldstein, "On the use of evaluative category scales in telecommunications," in *Proc. Human Factors Telecommun.*, Melbourne, VIC, Australia, Jan. 1995, pp. 253–260.

[37] Q. Huynh-Thu, M.-N. Garcia, F. Speranza, P. Corriveau, and A. Raake, "Study of rating scales for subjective quality assessment of high-definition video," *IEEE Trans. Broadcast.*, vol. 57, no. 1, pp. 1–14, Mar. 2011.

[38] R. Aldridge, J. Davidoff, M. Ghanbari, D. Hands, and D. Pearson, "Measurement of scene-dependent quality variations in digitally coded television pictures," *IEE Proc.-Vis., Image, Signal Process.*, vol. 142, no. 3, pp. 149–154, Jun. 1995.

[39] F. Zhang and D. R. Bull, "A parametric framework for video compression using region-based texture models," *IEEE J. Sel. Topics Signal Process.*, vol. 5, no. 7, pp. 1378–1392, Nov. 2011.

[40] D. C. Howell, *Statistical Methods for Psychology*, 8th ed. Boston, MA, USA: Cengage Learning, Jan. 2012.

[41] I. E. Sobel, "Camera models and machine perception," Ph.D. dissertation, Dept. Elect. Eng., Stanford Univ., Stanford, CA, USA, 1970.

[42] N. Kingsbury, "Complex wavelets for shift invariant analysis and filtering of signals," *Appl. Comput. Harmonics Anal.*, vol. 10, no. 3, pp. 234–253, 2001.

Felix Mercer Moss received the B.Sc. degree from University of Leeds, Leeds, U.K., in 2008, and the M.Res. and Ph.D. degrees from the Department of Experimental Psychology and the Department of Computer Science, University of Bristol, Bristol, U.K., in 2009 and 2013, respectively.

He is a Research Assistant with the Visual Information Laboratory, Department of Electrical and Electronic Engineering, University of Bristol, where he is involved in projects related to video quality assessment. His research interests include subjective video quality assessment, image evaluation, individual differences in eye movement control, and decision making.

Ke Wang received the B.Eng. degree in automation from Xi'an Jiaotong University, Xi'an, China, in 2013. She is currently working toward the M.Sc. degree in image and video communications and signal processing with University of Bristol, Bristol, U.K.

Her research interests include video quality assessment and video compression.

Fan Zhang (M'12) received the B.Sc. and M.Sc. degrees from Shanghai Jiao Tong University, Shanghai, China, and the Ph.D. degree from University of Bristol, Bristol, U.K.

He is a Research Assistant with the Visual Information Laboratory, Department of Electrical and Electronic Engineering, University of Bristol, where he is involved in projects related to parametric video coding and immersive technology. His research interests include perceptual video compression, video metrics, texture synthesis, subjective quality assessment, and HDR formation and compression.

Roland Baddeley received the B.Sc. degree from University of Sussex, Brighton, U.K., and the Ph.D. degree from University of Stirling, Stirling, U.K.

His previous roles include positions with the Department of Cambridge Physiology, the Department of Oxford Psychology, the Department of Physiology, and the Department of Sussex Psychology. He is currently a Reader in computational neuroscience with the School of Experimental Psychology, University of Bristol, Bristol, U.K. He has authored over 50 academic papers. His research interests include the statistics on natural images, eye movement control, camouflage, decision making, and the perception of texture.

David R. Bull (M'94–SM'07–F'12) received the B.Sc. degree from the University of Exeter, Exeter, U.K., in 1980; the M.Sc. degree from University of Manchester, Manchester, U.K., in 1983; and the Ph.D. degree from the University of Cardiff, Cardiff, U.K., in 1988.

He has previously been a Systems Engineer with Rolls Royce, Bristol, U.K., and a Lecturer with the University of Wales, Cardiff, U.K. He joined the University of Bristol in 1993, and is currently its Chair of Signal Processing and Director of its Bristol Vision Institute. In 2001, he co-founded a university spin-off company, ProVision Communication Technologies Ltd., specializing in wireless video technology. He has authored over 450 papers on the topics of image and video communications and analysis for wireless, Internet, and broadcast applications, together with numerous patents, several of which have been exploited commercially. He has received two IEE Premium Awards for his work. He is the author of three books and has delivered numerous invited/keynote lectures and tutorials.

Dr. Bull is a fellow of the Institution of Engineering and Technology.